Requirements for the domain model of environmental computational spreadsheets

Martine de Vos*
Computer Science
Network Institute
VU University Amsterdam
The Netherlands
martine.de.vos@vu.nl

Jan Wielemaker
Computer Science
Network Institute
VU University Amsterdam
The Netherlands
j.wielemaker@vu.nl

Guus Schreiber
Computer Science
Network Institute
VU University Amsterdam
The Netherlands
guus.schreiber@vu.nl

Jan Top
Wageningen University and
Research Centre
Food and Biobased Research
Wageningen The Netherlands
j.l.top@vu.nl

1. INTRODUCTION

Environmental computational models are considered essential tools in supporting environmental decision making by exploring the consequences of alternative policies or management scenarios [1, 2]. Environmental computational models are mainly developed and used by domain scientists and typically implemented as spreadsheets, Fortran programs or in MatLab. These domain scientists have a domain model in their minds, i.e., a knowledge level [3] model containing the important concepts in their domain and corresponding definitions and relations. In the model development process they inevitably make choices about which entities and processes they should include to describe their study area, and how these should be translated and implemented in the computational model. In this way their domain model is implicitly included in the computational model [4].

The domain model is essential for understanding the meaning and context of the results and insights generated with these models. As a consequence, it is hard to understand and reuse the domain knowledge in environmental computational models by other people than the original developers. The focus of this research is on environmental computational models that are implemented as spreadsheets, from now on called "environmental computational spreadsheets". The ultimate goal is to develop a set of semi-automatic methods for supporting the explication of the underlying domain model of environmental computational spreadsheets.

An important first step is to determine how such a domain model can be adequately described. We consider an adequate desription as a description that agrees with the views of the original developer(s) and can be understood and used by peers and stakeholders. In two case studies on the same dataset we discovered specific and concrete requirements that a domain model of a environmental computational spreadsheet should meet. In this paper we discuss these requirements, which we formulate as challenges, as there are currently no tools or methods available to fulfill these requirements.

KCAP '15 Palisades, NY, USA

2. DATA SET

Our data set is an existing scientific spreadsheet model for energy policy analysis [1]. Model calculations as well as input and supporting data are represented in several interconnected Excel workbooks. The main calculation workbook contains 39 spreadsheets, of which 22 are actually used in the calculation of results. These spreadsheets contain tables with both text, numbers and formulas, and contain a total of 79,059 cells with content. The spreadsheet model and analyses are described in a research report [5].

3. COMBINING DESCRIPTIVE AND COMPUTATIONAL KNOWLEDGE

3.1 Case study

In the first case study [6] we explore to which extent the domain model of an environmental computational spreadsheet can be made explicit. We manually analyze both the content and the design of the tables in a small set of spreadsheets from the data set. We analyze the various layout patterns in the spreadsheet and determine to what extent these patterns provide insight in the semantics of the content. Next, we semantically characterize the terms of the spreadsheet as instances of concepts of an existing ontology, i.e., the OM Ontology for Units of Measure and related concepts [7]. Four main concepts from the OM ontology were recognized in the spreadsheets: *Phenomenon, Quantity,Unit of Measure and Measure* (Figure 1). Finally, we analyze the

[1]Edesign,`http://www.pbl.nl/e-design/`

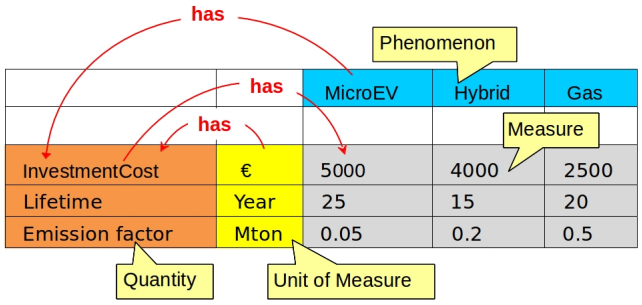

Figure 1: Example, in outline, of the color markup of the main used concepts and their relations in one of the spreadsheet tables [6]

formulas in the spreadsheet, and determine how the recognized OM concepts are connected through these formulas.

During the analysis process we observe the consecutive steps needed to recognize the semantics and record these in heuristics, for example, *The body of a spreadsheet table contains only Measures*, and *The headers of a spreadsheet table contain either Quantities or Phenomena*. The results of our manual analysis provide us with the domain concepts and relations in the spreadsheets, which we formally describe in an ontology (Figure 2). Subsequently, we interview the original developers of the spreadsheets to compare our findings with their views.

In this case study we found that the formulas in the spreadsheets contain implicit knowledge on the underlying semantics, as these represent connections between concepts in the domain model. With the semantics of part of the terms in the spreadsheet already known, the semantics of missing or ambiguous the terms could be deduced by combining knowledge from the formulas and the OM ontology.

We could not find any inconsistencies between our constructed ontology and the developers' views. From the interview it was clear that the developers were primarily focused on the calculation workflow, and showed limited interest in the ontology. This difference in focus made it difficult to perform an actual evaluation of our ontology.

3.2 Challenge

The domain knowledge included in spreadsheets can be viewed from different perspectives, i.e., a computational and a descriptive perspective. Domain scientists may see environmental computational spreadsheets mainly as instruments to perform simulation studies, and therefore focus primarily on the computational aspects. As they underestimate, or do not understand, the role of environmental computational spreadsheets in communicating scientific knowledge, they are less interested in the descriptive aspects.

In our opinion, the two perspectives are complementary and equally important, and, as shown in our case study, interconnected. The challenge is to create a reconstruction of the domain model that combines both perspectives, and to do this in an (semi)automated way.

Such a representation would provide a complete picture of the spreadsheet model. Furthermore, since it matches the view of the original developers, it will facilitate the evaluation procedure.

3.3 Possible approaches

Creating a "combined" domain model from a set of environmental computational spreadsheets would require the following actions:

1. deriving a description of the domain knowledge, i.e., concepts and relations, included in the spreadsheets
2. deriving a description of the calculation workflow
3. combining both descriptions in a meaningful way

The first step we performed manually in our case study (Figure 2), but we see several possiblities for automating this process. Domain ontologies could be used to automatically annotate domain concepts in spreadsheet tables. The information on table design, recorded as the heuristics in the case study, could be implemented in algorithms that inform the automatic annotation process.

The calculation workflow of a set of spreadsheets should provide insight in how results are calculated. Simply parsing the formulas in the spreadsheets results in a cell dependency graph that may contain thousands of nodes and edges, which is way beyond the limits of human visual comprehension. In recent work [8] we propose an approach for semi-automatically deriving the calculation workflow from the cell dependency graph, by aggregating it based on the analysis of the formula syntax and application of heuristics. Results from three case studies show that our constructed calculation workflows approximate the ground truth workflows both in size and content.

Our case study provides some useful pointers for performing the last step, i.e., linking the two descriptions. The use of OM, or a comparable meta-level ontology like QUDT [2], may play an in important role in this process. Characterizing terms in the spreadsheet as *Phenomena* and *Quantities* provides information on their roles in the structural domain model, as *Phenomena* can be considered domain concepts and *Quantities* the quantitative properties of these concepts. At the same time, we observed that the spreadsheet formulas only refer to *Quantities*. The *Quantities* in the tables could thus be linked to the variables in the calculation workflow, for example like in Figure 3

An important observation from our case study is that the different methods that we used for spreadsheet analysis and interpretation can inform each other. As such, we think that combining these methods in an iterative process would be a suitable and promising approach to explicate the knowledge that is implicitly included in spreadsheets. We also think that an iterative design could facilitate automating the interpretation process.

3.4 Related work

Several studies in different fields of computer science provide useful approaches that can be applied to the various stages of automatic construction of the domain model of environmental computational spreadsheets.

There are a few studies that automatically extract, analyze and visualize information in spreadsheets to support user understanding. Hermans and colleagues [9] automatically extracted information from spreadsheets, and used a library of common spreadsheet design patterns to transform it into class diagrams. They also analyzed cell dependencies

[2]QUDT, `http://www.qudt.org`

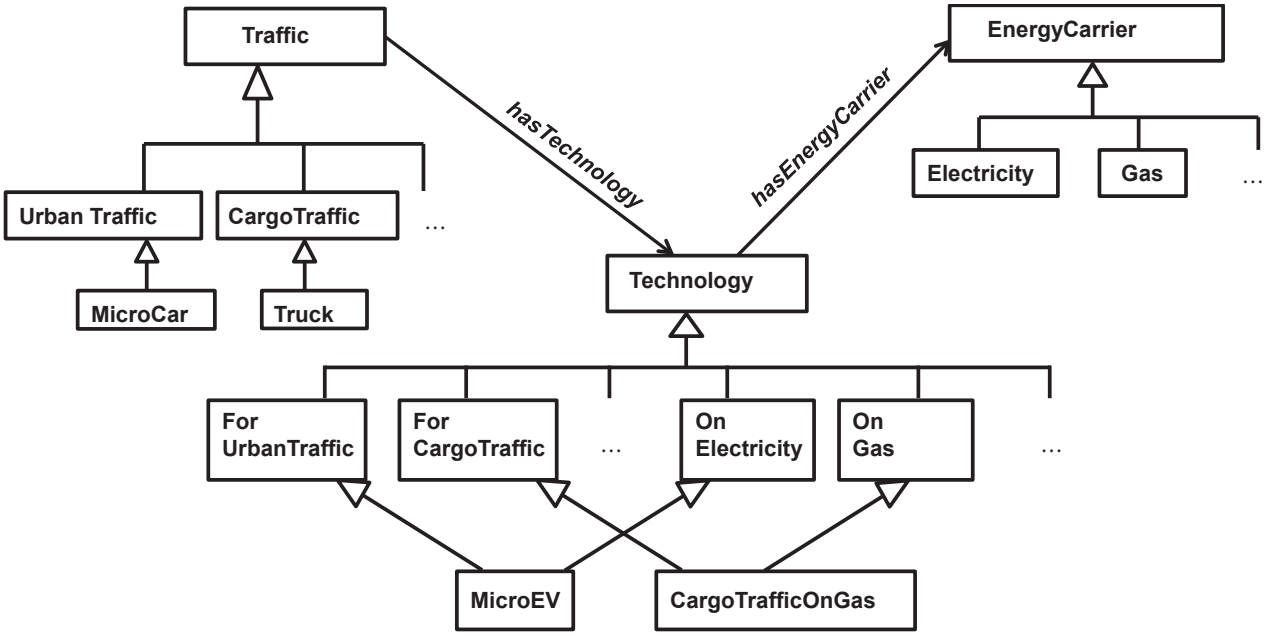

Figure 2: Example of manually constructed domain model [6]

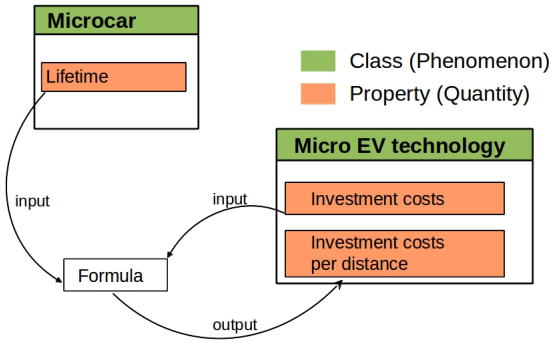

Figure 3: Example of combining both descriptive and computational domain knowledge in one model

in spreadsheets and and apply visual abstraction techniques to present these as leveld dataflow diagrams [10]. Clermont and colleagues [11] developed a toolkit that aggregates the cell dependencies in spreadsheets based on both formula syntax and data flow, and visualizes the results in graph-based presentations.

Mulwad and colleagues [12] used external ontologies and vocabularies to interpret knowledge from tabular data.

And especially relevant to the abovementioned challenge is the work of Borst and colleagues [13], who developed an ontology collection to describe the knowledge in dynamic physical systems. This collection contains different types of ontologies to describe, e.g., technical components, physical processes, and specific domain knowledge.

4. CREATING LINKS BETWEEN PUBLICATIONS AND SPREADSHEETS

4.1 Case study

In the second case study [14] we investigate to what extent links between environmental computational spreadsheets and the corresponding publications can be made explicit. These spreadsheets and the performed analyses are typically described in papers or reports, which provide readers with an explanation of the underlying concepts and an interpretation of the results. In practice, these publication serve as the single source of information on the underlying research project. However, it would be desirable if it was linked to key elements of the relevant computational model. In this way, the publication can provide insight into the model structure and calculation of results, and therefore provide a complete picture of the underlying research.

We automatically determine frequent terms in the spreadsheets and the corresponding publication, and compared both sets of terms (Table 1,2). Furthermore, we manually reconstruct calculation procedures from the storyline in the

Table 1: Top ten terms found in both spreadsheets and publication (according to spreadsheet rank)

term	publication tf-idf	spreadsheet count
pj *(peta joule)*	9.8	3667
pessimistic	5.6	1057
optimistic	7.5	1056
twh *(tera watt hour)*	10.8	828
heat	10.3	758
mton *(mega ton)*	27.3	717
biomass	40.3	667
km *(kilometer)*	3.5	495
co2	38.5	366
natural gas	6.1	365

Table 2: Number of terms per concept in spreadsheet and publication

Concept	# terms		
	spreadsheet	*publication*	*overlap*
Technology	78	31	27
Sector	37	26	19
Supply *(stock)*	24	13	11
Biomass	17	2	2

publication text and investigate to what extent these agree with the calculation procedures included in the spreadsheets (Figure 4).

The results of the term analysis showed that the publication and spreadsheets use the same concepts. But the publication typically focuses on the super and aggregate classes, while in the spreadsheets the low-level classes of the same concepts are more frequent. For example, the publication may only discuss *biomass* in general, while in the spreadsheets, multiple types of biomass are distinguished, like *manure*, *starch* and *wood*. We also found that the calculation procedure of model results, as described in the publication gives a correct, but incomplete and very general outline of the workflow included in the spreadsheets. The publication describes many aggregate or abstract variables which are not found as such in the spreadsheets, while the component variables from the spreadsheets are not present in the chapter. For example, the aggregate variable *total energy demand* is only found in the publication, while its components, like *energy demand from traffic* and *energy demand from industry*, are only found in the spreadsheets.

4.2 Challenge

As mentioned above, written publications are often the only information source for users and stakeholders on a research project. Constructing the domain model of environmental computational spreadsheets by itself may therefore not be sufficient to provide them with the domain knowledge included in these spreadsheets. There should also be explicit links with publications.

Results of the case study showed that the publication and spreadsheets use the same concepts, so linking the two items seems appropriate. The difference in abstraction level makes it difficult, however, to create direct links between elements in the report and the spreadsheets.

The challenge is to construct a domain model, that comprises several levels of conceptual abstraction, so it can serve as a hub between the publication and spreadsheets. Ideally, such a domain model would contain knowledge on both the concepts and computations (see above), as both aspects are described in the publication as well.

4.3 Possible approach

Using only terms from the spreadsheets for the domain model would not be sufficient, as these contain mainly low-level classes of domain concepts. An external domain ontology or vocabulary is needed to retrieve the higher levels of abstraction of these concepts that are needed to link these to concepts in the publication.

Spreadsheet developers usually group semantically related spreadsheet cells together, and use layout features to distinguish these groups [15, 16]. In the first case study it was observed that spreadsheet terms in the same group are related to the same high-level domain concept. As such, information on the design of the spreadsheet tables may provide a starting point to automatically retrieve high-level concepts from an external domain ontology or vocabulary

An approach to including both the concepts and computations of spreadsheets in the domain model is discussed in the first challenge.

4.4 Related work

Our work may benefit from the various approaches that are available for the annotation of scientific data sets. Several approaches have been developed to manually describe or annotate tabular data with concepts from external domain ontologies, e.g., Anzo suite [3], Rightfield [17] and Rosanne [7]. Annotation of concepts in text documents are made both manually and automatically, e.g., automatically connecting biomedical documents to terms from the Gene Ontology [18] and semi- automatic annotation of geo-spatial datasets with metadata provided by international guidelines from INSPIRE [19].

5. CONCLUSION

In this paper we identified two specific and concrete requirements that the domain model of an environmental computational spreadsheet should meet, i.e., 1) it should combine both descriptive and computational knowledge, and 2) it should comprises several levels of conceptual abstraction.

Currently, no existing methods or tools are available to automatically construct the domain model of environmental computational spreadsheets. However, we see several opportunities that could facilitate this task. Domain ontologies may be used to annotate domain knowledge terms in the spreadsheets and to add additional levels of conceptual abstraction to the domain model. Heuristics on the layout and structure of spreadsheet tables may be used to inform the annotation and interpretation process. And, a meta-level ontology may be used to characterize the quantities in spreadsheets both as part of the structural domain knowledge, i.e., as quantitative properties of phenomena, and as part of the calculation workflow, i.e., as variables. Subsequently, these quantities may serve as a link between the descriptive and computational knowledge in the domain model.

Overcoming these limitations would result in domain models that are understandable and accessible for domain scien-

[3]Anzo, `http://www.cambridgesemantics.com/)`

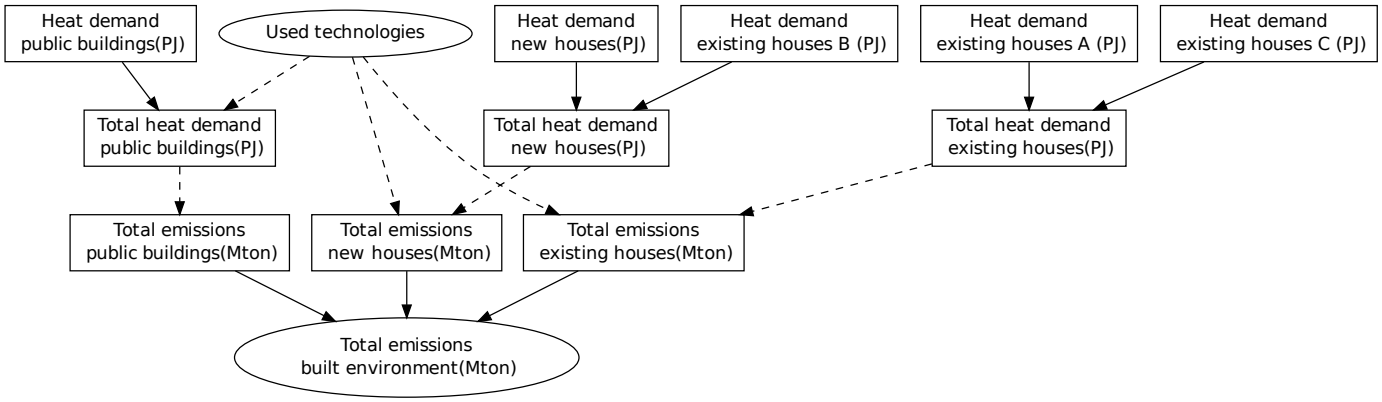

Figure 4: Reconstructed, simplifed calculation workflow of a final result spreadsheet cell. Record shape variables were only present in the spreadsheets, ellipse shape variables were also found in the publication[14]

tists. Ideally, these domain models will facilitate the communication of scientific knowledge and contribute to the construction of a shared knowledge base among environmental scientists.

The suggested approach of iteratively combining different methods of spreadsheet analysis and interpretation is innovative and promising. Ideally, this approach could facilitate automatic interpretation of (implicit) knowledge included in environmental computational spreadsheets. This would be benefecial to computer science.

Acknowledgments

We wish to thank PBL researchers Jan Ros en Jeroen Peters for providing us with their model and data, and our colleagues Willem van Hage and Bob Wielinga for their useful comments. This publication was supported by the Dutch national program COMMIT.

6. REFERENCES

[1] Jakeman, a., Letcher, R., Norton, J.: Ten iterative steps in development and evaluation of environmental models. Environmental Modelling & Software **21**(5) (May 2006) 602–614

[2] Schmolke, A., Thorbek, P., DeAngelis, D.L., Grimm, V.: Ecological models supporting environmental decision making: a strategy for the future. Trends in ecology & evolution **25**(8) (August 2010) 479–86

[3] Newell, A.: The knowledge level. Artificial Intelligence **18**(1) (January 1982) 87–127

[4] De Vos, M.G.: Interpreting environmental computational spreadsheets. In Groth, P., Noy, N., eds.: Proceedings of the Doctoral Consortium at the 13th International Semantic Web Conference (ISWC 2014), Riva del Garda, Italy (2014)

[5] PBL Planbureau voor de Leefomgeving, Energieonderzoek Centrum Nederland: Naar een schone economie in 2050: routes verkend. Technical report, Netherlands Environmental Assesment Agency, Den Haag (2011)

[6] De Vos, M., Van Hage, W.R., Ros, J., Schreiber, A.: Reconstructing Semantics of Scientific Models : a Case Study. In: Proceedings of the OEDW workshop on Ontology engineering in a data driven world, EKAW 2012, Galway, Ireland (2012)

[7] Rijgersberg, H., Wigham, M., Top, J.: How semantics can improve engineering processes: A case of units of measure and quantities. Advanced Engineering Informatics **25**(2) (April 2011) 276–287

[8] De Vos, M.G., Wielemaker, J., Wielinga, B., Schreiber, G., Top, J.: A methodology for constructing the calculation model of scientific spreadsheets. In: Proceedings of the 8th International Conference on Knowledge Capture. (2015)

[9] Hermans, F., Pinzger, M., Deursen, A.V.: Automatically Extracting Class Diagrams from Spreadsheets. In: 24th European Conference on Object-Oriented Programming (ECOOP),Lecture Notes in Computer Science,, Springer-Verlag (2010) 52–75

[10] Hermans, F., Pinzger, M., Deursen, A.V.: Supporting Professional Spreadsheet Users by Generating Leveled Dataflow Diagrams. In: Proceedings of the 33rd International Conference on Software Engineering., ACM (2011)

[11] Clermont, M.: A Toolkit for Scalable Spreadsheet Visualization. In: Proceedings of EuSpRIG 2004 Conference, European Spreadsheet Risks Interest Group (2004) 1–12

[12] Mulwad, V., Finin, T., Joshi, A.: A Domain Independent Framework for Extracting Linked Semantic Data from Tables. In: Search Computing. Springer Berlin Heidelberg (2012) 16–33

[13] Borst, P., Akkermans, H., Top, J.: Engineering Ontologies. International Journal of Human-Computer Studies (2007) 365–406

[14] De Vos, M., van Hage, W.R., Wielemaker, J., Schreiber, A.: Knowledge Representation in Scientific Models and their Publications : a Case Study. In: Proceedings of the 7th International Conference on Knowledge Capture., Banff, Canada (2013) 1–2

[15] Mittermeir, R., Clermont, M.: Finding High-Level Structures in Spreadsheet Programs. In: Proceedings of the 9th Working Conference on Reverse Engineering, Richmond,VA,USA (2002) 221–232

[16] Hipfl, S.: Using Layout Information for Spreadsheet Visualization. In: Proceedings of the European Spreadsheet Risks Interest Group 5th Annual Conference, Klagenfurt,Austria (2004)

[17] Wolstencroft, K., Owen, S., Horridge, M., Krebs, O., Mueller, W., Snoep, J.L., du Preez, F., Goble, C.: RightField: embedding ontology annotation in spreadsheets. Bioinformatics (Oxford, England) **27**(14) (July 2011) 2021–2

[18] Smith, T.C., Cleary, J.G.: Automatically linking MEDLINE abstracts to the Gene Ontology. In: Proc. ISMB 2003 BioLINK Text Data Mining SIG. (2003) 1–4

[19] Macário, C.G.N., de Sousa, S.R., Medeiros, C.B.: Annotating geospatial data based on its semantics. In: Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - GIS '09, New York, New York, USA, ACM Press (2009) 81